DOCUMENT RESUME

ED 110 503 TM 004 776

TITLE	Computer Assisted Assembly of Tests at Educational Testing Service.
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	[Oct 74]
NOTE	14p.; Paper presented at the Computer Assisted Test Construction Conference (San Diego, California, October 14-16, 1974); For a related document, see TM 004 778
EDRS PRICE	MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS	*Classification; Computer Oriented Programs; *Computer Programs; Data Bases; Data Processing; *Item Banks; *Test Construction; Tests; *Test Selection

ABSTRACT

Two basic requirements for the successful initiation of a program for test assembly are the development of detailed item content classification systems and the delineation of the professional judgements made in building a test from a pool of items to detailed content, ability, and statistical specifications in terms precise enough to be translated into computer programs. A guiding principle has been that the computer is to serve the professional staff, not supersede them. Every test assembled is subject to professional review; revisions can be requested; and the items selected are held "in limbo" until the test has been accepted. The systems development turned out to be much more sophisticated than originally expected. It is undoubtedly more complex than would be required for most testing situations since it is designed to handle a wide variety of tests in each subject area and the numerous constraints that are imposed on item selection by the rigorous test specifications and need for parallelism of forms that is basic to some of Educational Testing Service's (ETS) national programs like the College Board tests. In addition to a unique identification number and the classification for each item, the computerized file contains a complete history of the item's uses, up to five sets of statistics, codes for the security level and present activity status of the item, and up to 15 12-letter key words--which in the case of some verbal items constitute the complete items. (Author/BJG)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED110503

COMPUTER ASSISTED ASSEMBLY OF TESTS

at Educational Testing Service

A paper presented at

The Computer Assisted Test Construction Conference

San Diego, California

October 1974

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM 004 776

Educational Testing Service
Princeton, New Jersey

Computer Assisted Assembly of Tests at Educational Testing Service

Toward the end of 1966, the Test Development and Data Processing Divisions at ETS began working together on the development of an operational system for assembly of tests by computer. This work was preceded by almost two years of experimentation and planning, including a simplified pilot assembly of a mathematics achievement test and the development of a mathematical model for the simulation of the logical processes involved in the selection of reading comprehension passages, to demonstrate feasibility.

The Verbal Aptitude and Mathematics areas were selected for the first phase because the item pools were largest, the number of tests to be assembled each year were greatest, and the SAT test-specifications most detailed and rigorous. Any system capable of assembling these tests could be easily adapted for other subject areas. Aptitude Tests have been routinely assembled by computer since 1968. We have also experimentally assembled mathematics achievement tests and have pilot tested the adaptation of the system to social studies by assembly of American History Tests. However, we have not pressed the expansion of the system to other areas because of some problems in developing classification systems and the realization that the present procedure is not cost effective in areas with relatively small item pools and infrequent test assemblies. Our present computerized item files include about 15,000 Verbal items, 12,000 Mathematics items and several hundred American History items. The flexibility of the system has been demonstrated by the adaptation and simplification of the program packages for a Russian Language pool for the Defense Language Institute.

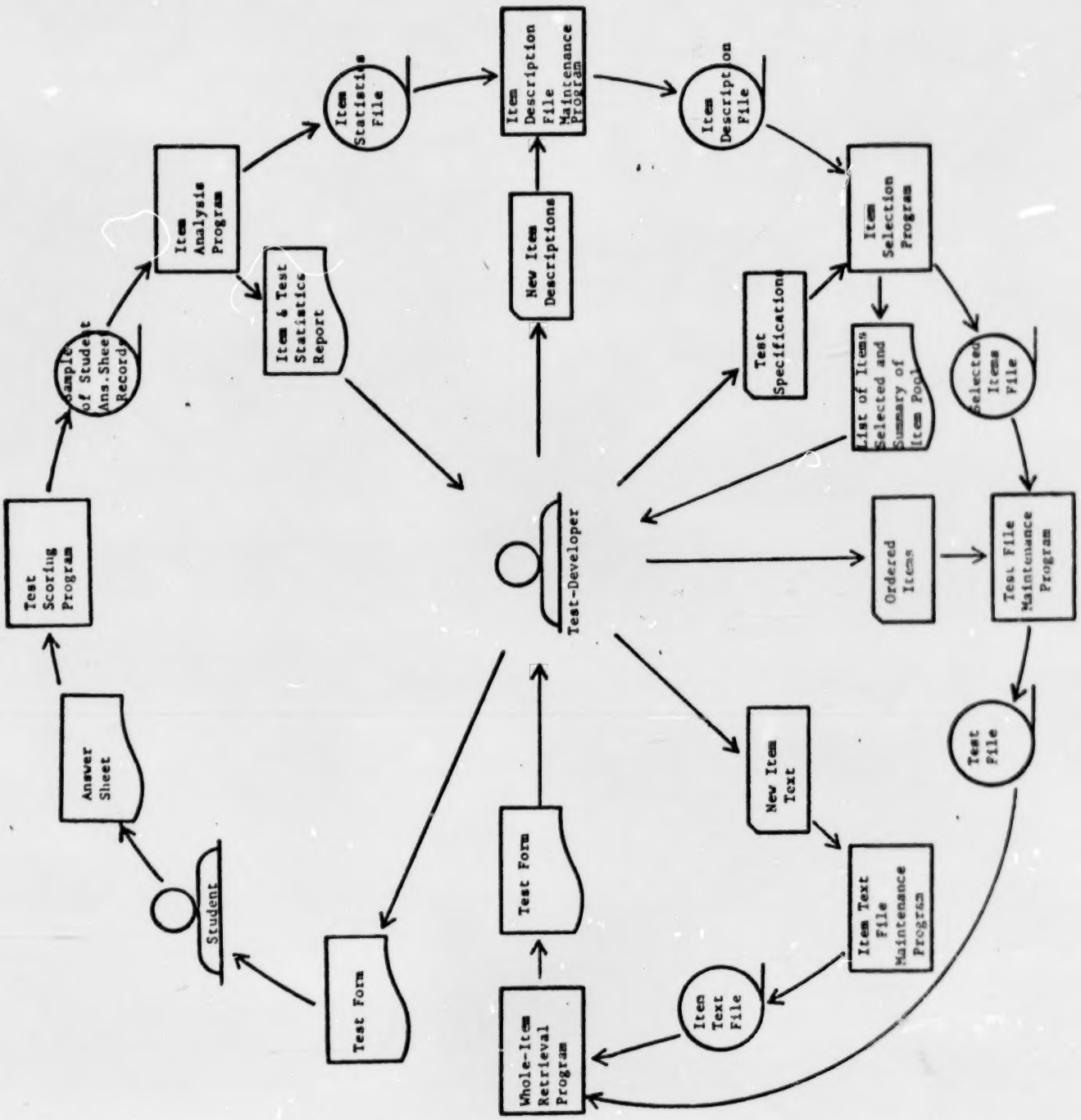
It was agreed from the outset that there were two basic requirements for the successful initiation of any such program - the development of detailed item content classification systems and the delineation of the professional judgments made in building a test from a pool of items to detailed content, ability, and statistical specifications in terms precise enough to be translated into computer programs. Our guiding principle has been that the computer is to serve the professional staff, not supersede them. Every test assembled

is subject to professional review; revisions can be requested; and the items selected are held "in limbo" until the test has been accepted. The accompanying figure shows the interaction between the test developer and the computer.

The systems development turned out to be much more sophisticated than we had originally expected. It is undoubtedly more complex than would be required for most testing situations since it is designed to handle a wide variety of tests in each subject area and the numerous constraints that are imposed on item selection by the rigorous test specifications and need for parallelism of forms that is basic to some of ETS's national programs like the College Board tests. In addition to a unique identification number and the classification for each item, the computerized file contains a complete history of the item's uses, up to five sets of statistics, codes for the security level and present activity status of the item, and up to fifteen twelve-letter key words - which in the case of some verbal items, constitute the complete items.

The basic computer programs in the system are a file maintenance package and an item selection package. The comprehensive inventories produced almost as a by-product of the selection package have turned out to be one of the most valuable features of the system. By providing a summary of items available for any category by difficulty level and discrimination index, they are an invaluable guide to where attention should be focused in item writing to keep a balanced pool. It is also possible to determine whether our item pools are adequate to meet requests for special tests and to assemble them quickly. An example of an inventory print-out for one type of verbal items is attached.

The selection program is not a simple item retrieval system; rather it is designed to fit multidimensional test specifications. Control of the test assembly process remains in the hands of the test developer because it is he who determines the specifications and reviews the product. When tests are requested from the computer, the statistical specifications provided must include acceptable ranges of means and standard deviations of difficulty and discrimination indices, and item selection is guided by designating the number of items need from each major category and for each minor category digit,


Item Statistic File
Item Description File Maintenance Program
Item Description File
Item Analysis Program
Item & Test Statistics Report
New Item Descriptions
Item Selection Program
Test Specifications
List of Items Selected and Summary of Item Pool
Selected Items File
Test Scoring Program
Test-Developer
Ordered Items
Test File Maintenance Program
Answer Sheet
Student
Test Form
Test Form
New Item Text
Item Text File Maintenance Program
Item Text File
Whole-Item Retrieval Program
Test File

ANTONYMS - SAMPLE INVENTORY Free Use Items (SAT)

1168 ITEMS ARE IN THE REDUCED POOL

MEAN DELTA = 12.31
DELTA STD DEV = 2.7357

MEAN R-BIS = 0.426

DELTA	FREQUENCY
0.0 - 4.9	7
5.0 - 5.9	26
6.0 - 6.9	15
7.0 - 7.9	30
8.0 - 8.9	73
9.0 - 9.9	75
10.0 - 10.9	119
11.0 - 11.9	121
12.0 - 12.9	181
13.0 - 13.9	199
14.0 - 14.9	114
15.0 - 15.9	126
16.0 - 16.9	61
17.0 - 17.9	10
18.0 - 18.9	9
19.0 - 19.9	2
20.0 - 26.0	

R-BIS	FREQUENCY
0.0 - 0.09	
0.10 - 0.19	
0.20 - 0.29	176
0.30 - 0.39	306
0.40 - 0.49	310
0.50 - 0.59	284
0.60 - 0.69	92
0.70 - 0.79	
0.80 - 0.89	
0.90 - 0.99	

DELTA	R-BIS 0.0 - 0.09	0.10 - 0.19	0.20 - 0.29	0.30 - 0.39	0.40 - 0.49	0.50 - 0.59	0.60 - 0.69	0.70 - 0.79	0.80 - 0.89	0.90 - 0.99	MEAN R-BIS
0.0 - 4.9			2	1	4						.467
5.0 - 5.9			9	9	8						.445
6.0 - 6.9			3	4	5	3					.481
7.0 - 7.9			7	6	10	7					.501
8.0 - 8.9		1	15	15	25	17					.496
9.0 - 9.9		2	13	15	32	13					.492
10.0 - 10.9		2	18	37	43	19					.493
11.0 - 11.9		1	19	34	52	15					.488
12.0 - 12.9		6	39	70	54	12					.459
13.0 - 13.9		20	56	81	39	3					.420
14.0 - 14.9		16	60	31	5	2					.370
15.0 - 15.9		78	39	3	5	1					.292
16.0 - 16.9		35	21	3	2						.299
17.0 - 17.9		9		1							.266
18.0 - 18.9		5	4								.279
19.0 - 19.9		1	1								.295
20.0 - 26.0											
MEAN DELTA		15.2	12.9	11.9	11.0	10.3					

ANTONYM INVENTORY (Continued)

SUMMARY OF CONTENT

LEVEL	DIGIT								
	1	2	3	4	5	6	7	8	9
0				1	1	1168	1168	1168	1168
1	282	604	839	337					
2	302	564	326	308	1				
3	214		3	510					
4	352			12					
5	18								
6					1163				
7									
8					3				
9									

MINOR CONTENT MEAN DELTA

LEVEL	DIGIT								
	1	2	3	4	5	6	7	8	9
0				9.10	13.30	12.31	12.31	12.31	12.31
1	12.80	12.54	12.38	11.80					
2	11.77	12.06	12.13	12.39	6.90				
3	11.55		10.56	12.58					
4	12.84			12.90					
5	12.07								
6					12.30				
7									
8					15.73				
9									

MINOR CONTENT MEAN R-BIS

LEVEL	DIGIT								
	1	2	3	4	5	6	7	8	9
0				.429	.349	.426	.426	.426	.426
1	.418	.418	.423	.439					
2	.435	.433	.433	.420	.309				
3	.424		.463	.420					
4	.424			.414					
5	.428								
6					.426				
7									
8					.256				
9									

indicating acceptable limits by requesting exactly so many items, at least so many items, or at most so many items. The computer surveys the entire pool, produces a reduced item pool of all those items that could be eligible for the particular test, and then begins a selection process by assigning a priority number to each item according to a complex procedure.

The highest priority item is then selected. This means that the entire pool of items is searched every time an item is selected. Therefore, it can be stated that every item is picked specifically because it best meets the specifications. After each item is selected, new priority numbers are assigned to a further reduced pool. In the verbal assembly package, an array of key words used is added to and checked after each item is selected since specifications require that no substantive word be used more than once in a test. Controls are built-in through the priority number computation to insure leaving the pool as balanced as possible after a test is selected to maximize the number of parallel tests that can be assembled.

Items are classified (and it follows that tests can be assembled) on the following parameters:

- Delta - difficulty of the item as determined by item analysis. A distribution across 15 delta levels may be specified for the test assembly as well as test mean and standard deviation. Furthermore, the item selection program allows for delta scaling "on the fly" for compensation across testing programs (or grade levels).
- R-bis - discriminating power of item as determined by item analysis. Required test mean and standard deviation are specified for assembly.
- Content - a nine-digit classification scheme to describe item content can be accommodated by the system. Furthermore, items can be classified into broad major categories within an overall subject. The semantics of this classification scheme are independent of the program.
- Item Security/Item Activity - two codes to indicate the current use or status of the item.
- Key Words - up to a dozen 15-character words appearing in or

descriptive of the item.

There are three primary files associated with the CAAT System.

- The Item Description File contains: 1) item classification data, 2) up to 10 previous uses of the item (tests in which it has appeared), 3) up to five complete sets of item analysis statistics, and 4) key words.
- The Test Description File is a history of which tests have used which items. For each test, items are listed in the order they appeared within each test section. This provides the means of matching new Item Analysis data to the items in the Item Description File, since the CAAT item identification numbers (accession numbers) are not known by the Item Analysis program.
- The Selected Items File is a transient storage area for tests which have been selected by CAAT but not yet approved by the test developer. When final approval has been given, the test is removed from the Selected Items File, added to the test file and the corresponding Item Description records are updated to reflect the new use.

The basic content classification format allows for any number of four-letter major categories, and within each of these categories up to nine digits, each with ten levels - a staggering number of possible unique classifications. In Verbal Aptitude, the major categories are acronyms for item types such as ANAL for analogies, ANTM for antonyms or RC for reading comprehension; in Mathematics they are broad content areas such as ALGB for algebra. The digits are used for finer classifications of content or item type and for such things as grade level, and type of thinking or ability tested. A basic requirement is that a group developing and using such a system agree on the meaning of a classification and use it consistently, since the computer couldn't care less what the letters and numbers stand for.

In developing subject-matter classifications, we have tried to make them detailed enough to be useful for the assembly of unit and end-of-course tests

as well as broad survey-type tests such as the College Board Achievement Examinations. In developing ability classifications in each subject area, we have tried to reflect the objectives of education in the field in terms of cognitive levels or thought processes. Our present classifications may not be detailed enough for diagnostic testing within topics but there is room enough in the classification system to add this type of detail if it should become desirable. Although we have not yet tried it, we think we can use a combination of the content and ability dimensions to match individual school's behavioral objectives should we move to assembling tests on request for this purpose.

In planning the system, we developed a content classification format that can, hopefully, be generalized to all subject areas without limiting flexibility. The approach to content classification is completely different for mathematics and verbal items, with social studies more similar to mathematics, but the format looks the same to the computer and the same programs can be used for file development and maintenance and for making inventories of the pool. The item selection logic is fundamentally similar with some adaptation for special problems in each area. Attached to this paper are sample classifications of a mathematics, a verbal aptitude, and a social studies item. As mentioned before, for verbal items, the major classifications are item types with the digits used for finer distinctions and grade level. For mathematics and social studies, the major classifications are broad, horizontal subject-matter categories such as ARIT for Arithmetic or AMHY for American History. In mathematics the first digit, which we call a major-minor category, is used for up to ten large subdivisions of each major area such as algebraic operations or linear functions in Algebra. The second and third digits refer to the first digit and theoretically, if not practically, provide the possibility of 100 sub-classifications for each major-minor category and 1000 unique classifications within each major area. The fourth and fifth digits are not being used at this time in mathematics but are available for finer classifications. To our classifiers, "happiness is the fourth and fifth digit."

Digits 6-9 have been assigned the same meaning overall mathematic questions

as follows:

> Digit 6 controls with respect to curriculum and has three levels of traditional versus modern subject matter and symbolism.
>
> Digit 7 indicates grade level from 1 for primary to 9 for advanced college level.
>
> Digit 8 describes the nature of the setting to insure variety in appearance and tasks. The ten levels distinguish between abstract and concrete settings, with and without figures or graphs, etc.
>
> Digit 9 we call the ability process dimension. It is a hierarchial classification into six levels which is our pragmatic adaptation of the Bloom taxonomy to mathematical problems and abilities to be tested.

In social studies, the use and pairing of digits varies with the major category except that digits 7-9 are used across all areas for grade level, type of stimulus, and ability tested. All nine digits are used for History.

The task of classifying all the old items in our files and transcribing the data has been and continues to be a mammoth and incomplete one. However, new items are classified when written and put on tape immediately and a system has been developed for linking item analyses directly with the computerized item files.

The computer output from the test selection process is a listing by major category of item accession numbers with classification, delta (difficulty index) and r-biserial and any key words available. Summary statistics are provided for each category and the test as a whole along with an inventory of all items in the reduced pool from which the test was chosen.

As described above, the primary output of the present system is a list of item numbers. The actual items are retrieved manually, put into an appropriate order within the test, typed, and printed. The desirability of automating this

SAMPLE CLASSIFICATIONS

Mathematics

A	R	I	T	1	2	5	Ø	Ø	2	4	5	3

The above classification describes an arithmetic item testing powers/roots of odd/even integers. The content is modern; grade level, junior high school setting, abstract, numbers only; ability level, comprehension of a concept.

Verbal

A	N	T	M	3	2	1	3	6	Ø	Ø	Ø	Ø

The above classification describes an antonym item in which the stem has a science connotation. It tests a fine distinction; it is a single word (as opposed to a phrase); which is an adjective. The "grade level" is appropriate for SAT (11th-12th grade).

Social Studies

W	E	H	Y	4	2	3	3	2	5	4	4

The above classification describes an item in the history of Western Europe. Chronologically, it is 16th century; geographically, France; its subject matter is Governmental Institutions and the emphasis is Political-internal. The "grade level" is College specialized. The item is based on a quotation and the ability level being measured is application.

process has been recognized from the start of the development of CAAT, and studies of Whole Item Storage and Retrieval (W.I.S.A.R.) have been undertaken. The main difficulty has been in the handling of graphic items (pictures, charts, etc.). To achieve the quality of reproduction that has become standard at ETS would require very expensive hardware. Similarly, although not as severe a limitation, the type fonts available on a high-speed printer are not of the accustomed quality. ETS currently uses IBM Executive Selectric typewriters for final copy of examination booklets.

For demonstration purposes, a Whole Item Storage and Retrieval System has been implemented as an adjunct to the CAAT system. The programs were developed external to ETS and have been adapted with the help and kind cooperation of their author Dr. Willard Brown, Western Washington State College, Bellingham, Washington. Dr. Brown's program has been modified to type test copy using the CAAT Test File as its source of item numbers. IBM Selectric typewriter terminals may be used, however there is no graphic capability other than creative combination of type characters.

All of us involved in computerized assembly of tests at ETS are aware that what we have accomplished is a major step but only a first one. One continuing problem is the time and money for development of sufficiently large and diverse pools of pretested items in other subjects to have available questions at different levels of difficulty within each content and ability classification so that the computer can assemble tests to specifications with minimal compromises. Although the system contains the capacity to apply a linear adjustment to the difficulty indices if an item is to be used at a different level, we have not yet been able to do the studies necessary to ascertain appropriate adjustment factors. We continue to search for a satisfactory method of whole item storage and retrieval, looking toward both computerized test production and printing and interactive testing using computer terminals. Although hardware and technology for these already exist, we have not yet found it economically possible or the output of high enough quality for our present special requirements - particularly since so many of our

test questions use nonverbal stimuli such as figures, graphs, cartoons, etc.

What lies ahead for the future? If the use of the system is to be expanded, we must move to whole item storage with tests being produced as computer print-out or transmitted to remote terminals for individualized interactive testing at consoles. If, as seems likely, the need for large national administrations of tests for selection diminishes, there will be little impetus for expanding the present system of computer assisted test assembly. Instead, its greatest potential would seem to be its adaptation for the assembly of customized tests from item pools for guidance, placement, and evaluation, for individualized interactive testing, and for test administrations at computer terminals with immediate scoring and feedback. The technology for all of these already exists; what is still needed is time and money for development and the ability to make the procedures cost effective.